2015

Human Absorbable MicroRNA Prediction based on an Ensemble Manifold Ranking Model

Jiang Shu
University of Nebraska-Lincoln, jshu2@unl.edu

Kevin Chiang
University of Nebraska-Lincoln, t3-kchiang@unl.edu

Dongyu Zhao
University of Nebraska-Lincoln, dzhao3@unl.edu

Juan Cui
University of Nebraska-Lincoln, jcui@unl.edu

HHS Public Access

Author manuscript

Proceedings (IEEE Int Conf Bioinformatics Biomed). Author manuscript; available in PMC 2018 November 08.

Published in final edited form as:

Proceedings (IEEE Int Conf Bioinformatics Biomed). 2015 November ; 2015: 295–300. doi:10.1109/BIBM.2015.7359697.

Human Absorbable MicroRNA Prediction based on an Ensemble Manifold Ranking Model

Jiang Shu, **Kevin Chiang**, **Dongyu Zhao**, and **Juan Cui**

Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE, United States

Abstract

MicroRNAs, a class of short non-coding RNAs, are able to regulate more than half of human genes and affect many fundamental biological processes. It has been long considered synthesized endogenously until very recent discoveries showing that human can absorb exogenous microRNAs from dietary resources. This finding has raised a challenge scientific question: which exogenous microRNAs can be integrated into human circulation and possibly exert functions in human? Here we present a well-designed ensemble manifold ranking model for identifying human absorbable exogenous miRNAs from 14 common dietary species. Specifically, we have analyzed 4,910 dietary microRNAs with 1,120 features derived based on the microRNA sequence and structure. In total, 70 discriminative features were selected to characterize the circulating microRNAs in human and have been used to infer the possibility of a certain exogenous microRNA getting integrated into human circulation. Finally, 461 dietary microRNAs have been identified as transportable exogenous microRNAs. To assess the performance of our ensemble model, we have validated the top predictions through a milk-feeding study. In addition, 26 microRNAs from two virus species were predicted as transportable and have been validated in two external experiments. The results demonstrate the data-driven computational model is highly promising to study transportable microRNAs while bypassing the complex mechanistic details.

Keywords

Dietary microRNAs; viral miRNAs; cross-species transportable microRNAs; feature selection; ensemble manifold ranking model

I. Introduction

Mature microRNAs (miRNAs) are a class of short non-coding RNAs, which are typically 21–25 nucleotides long. In the past decade, numerous studies have shown that this type of small molecules can negatively regulate gene expression post-transcriptionally [1–3]. In most cases, miRNAs can bind to the target messenger RNAs (mRNAs) and prevent the protein products of corresponding mRNA by either inhibiting the translation process or promoting the mRNA decay [1,2]. More than 60% of human genes, at a conservative

Correspondence to: Juan Cui.

Jshu2@unl.edu, t3-kchiang@unl.edu, dzhao3@unl.edu, jcui@unl.edu.

estimate, can be targeted by 2,588 known human miRNAs [3]. The regulation of miRNAs significantly affects a number of fundamental biological processes and pathogenesis of human disease[4].

It was generally considered that miRNAs are synthesized endogenously within the individual. However, the latest study shows that human is able to absorb exogenous miRNAs from bovine milk [5] where the authors have successfully measured meaningful amounts of two cow's milk miRNAs, bta-miR-29b and −200c, in human blood. Moreover, the study also demonstrated the potential influences of two transferred cow miRNAs on human health through regulating human genes. Similarly, other experiments suggest that one rice's miRNA, osa-miR-168a, could also be transferred into the circulation of mammals [6]. These observations have raised a challenge question: which exogenous miRNAs can be absorbed and integrated into human circulation, then potentially play regulatory roles in human.

The cross-species transportation of miRNAs is an emerging research topic where the mechanism is largely unknown. Nevertheless, several studies uncovered two main forms of detected circulating miRNAs in human: either associated with exosomes (vesicles or microparticles) or bounded to Argonaute (AGO) proteins in RNAi silencing complex [7–11]. Either way, it requires a distinct binding pattern between the miRNA and another molecule. Therefore, the binding affinity of miRNA-protein is very likely to affect the possibility of cross-species transportation. Based on several such assumptions, we applied a data mining strategy to identify discriminative molecular features that may have an impact on the transportation, such as: nucleotide compositions on seed region, %G+C content of mature miRNA sequences [1,12,13] and many features generated from the secondary structure of precursor miRNAs including minimum free energy of the secondary structure and stem length [14–20]. As a result, 1,120 sequential and structural features that possibly affect the miRNA binding and transportation have been considered.

In this article, we present an ensemble manifold ranking model for identifying potential human absorbable exogenous miRNAs. 360 validated human circulating miRNAs from pervious finding were used to train the model to infer the most likely transportable exogenous miRNAs from 14 common food species. **To the best of our knowledge, this is the first study that aims to provide an efficient high-throughput computational screening for cross-species transportable miRNAs.**

II. Materials and Method

In this section, we provide a detailed description of our computational model, which includes the following sections: miRNA datasets, feature extraction and the ensemble manifold ranking model for prediction.

A. Datasets

Among 14 most common dietary species, we collected the sequences of 4,910 mature miRNAs and 4,387 corresponding stem-loop precursor miRNAs from the Dietary microRNA Database (DMD) developed by our group [21]. An independent validation set also includes sequence data from two virus species at *miRBase* [3]: Epstein-Barr virus and

Rhesus lymphocrypto-virus. TABLE I illustrates the detailed statistics of the miRNA data we used in this study.

In order to identify the potential exogenous miRNAs that can be integrated into human circulation, we retrieved 360 plasma miRNAs from Weber's study [22] and used them as a positive set to train the prediction model.

B. Feature extraction

As described above, we suspected that many sequential or structural features of miRNA likely differentiate the circulating miRNA against others. Therefore, we extracted 1,120 features [23–25] to assess their discriminative powers on the circulating miRNA prediction. Specifically, for each mature miRNA, a total of 1,102 features were generated including:

1. 1,031 features calculated based on following sequences:
    a. extend seed region sequence (first 8 nucleotides on 5' end of mature miRNA sequence);
    b. mature miRNA sequence;
    c. corresponding precursor stem-loop sequence.
2. 71 structural features identified based on the predicted secondary structure of precursor stem-loop sequence.

The detailed feature information can be found in TABLE II.

C. The ensemble manifold ranking model

Unlike a typical binary classification problem, prediction of possible transferable exogenous miRNAs is only conducted based on the known circulating miRNAs (positive only). Since it is quite possible that there are many miRNAs might be transportable to human circulation but have not been detected yet, it fails to define a negative set in our study. Thus, manifold ranking is employed here, which has been proven to be a powerful tool in the unaiy classification cases [26, 27].

1) Manifold ranking: Manifold ranking is a graph-based ranking algorithm that has been widely used in information retrieval and has shown to perform very well on a variety of datasets. It originally proposed as a personalized version of the PageRank algorithm [28], and were successfully applied on image data, textual data, and biological datasets [29–31].

The algorithm for Manifold Ranking is as follows:

a. Sort each sample based on max pairwise distance of its feature vector, and connect each sample until a connected graph is formed;
b. Form a distance matrix using the RBF kernel, assign 1 if there is an edge linking two samples, 0 otherwise;
c. Normalize the distance matrix using symmetric Laplacian normalization;

d. Spread a sample's ranking score to their neighbors according to the weighted network. Repeat this step until a stable state is achieved. This step contains a parameter α that specifies the relative contributions to the ranking scores from a sample's neighbors and the initial ranking scores;

e. Rank the nodes according to their ranking scores.

There are two parameters in the algorithm: σ for the RBF kernel setting and α for the weights of prior knowledge from the positive set. Some manifold ranking applications used the empirical parameters, σ as $\frac{\bar{d}}{3}$ ($\bar{d}$ is average distance among all samples) and α as 0.99 [31]. To avoid such arbitrary setting, our model conducts parameter search to ensure the best predictive performance is achieved.

One typical assessment of a ranking method is checking the percentage of the positive training data that is ranked among the top X% of all the training data. Generally the higher the percentage is for each fixed X, the better the trained ranking algorithm is.

1) Feature ranking generation: As mentioned above, we generated 1,120 features for each miRNA to characterize the human circulating miRNAs. However, it is very unlikely that every feature contributes on this recognition. Thus, we applied three different methods to evaluate the discriminative power of each feature:

a) F-score:

First, the F-Score Ranking was calculated as described by Chen and Lin [32]. It is a simple technique that measures the discrimination of two sets of real numbers. The F-score is defined as shown in the following equation, where $\bar{x}_1, \bar{x}_i^{(+)}$ and $\bar{x}_i^{(-)}$ are the average of the i^{th} feature of the whole, positive, and negative (unlabeled samples, in our case.) datasets respectively. The larger the F-score is, the more likely that the i^{th} feature is discriminative.

$$F(i) \equiv \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2}$$

b) Fisher's ratio

The second method is known as Fisher's ratio also attempts to rank features by linear discriminative power. It is defined as the difference in means squared over the difference in variance:

$$FR = \frac{(\bar{x}_i^{(+)} - \bar{x}_i^{(-)})^2}{v_i^{(+)} - v_i^{(-)}}$$

$v_I^{(+)}$and $v_i^{(-)}$are the variance of the i^{th} feature of the whole, positive and negative datasets.

c) Wilcoxon signed-rank test statistic W Wiloxon test statistic W is calculated as:

$$W_i = \sum_{j=1}^{N} [\mathrm{sgn}(x_{j,i}^{(+)} - x_{j,i}^{(-)}) \cdot R_j] \quad \mathrm{sgn}(t) = \begin{cases} -1 & t < 0 \\ 0 & t = 0 \\ 1 & t > 0 \end{cases}$$

Since each individual ranking presents a different evaluation of predictive power of each feature, we applied robust ranking aggregation method to integrate all the information into one final feature ranking for further feature selection. The aggregation is conducted by *RankAggreg* package in *R* [33]. It aggregates three independent rankings by using Cross-Entropy Monte Carlo algorithm and Spearman distance measurement. It should be noted here that, since there is no any pre-determined preference among three feature ranking methods, we assigned the same weight on each ranking lists.

2) Feature selection—While the traditional manifold ranking requires the preselected features, the ensemble manifold ranking model enables the selection of the discriminative features among the initial feature set and optimized parameters. At the end, along with the selected features and parameters, the model conducts a final manifold ranking with all positive set to identify the potential human absorbable exogenous miRNAs among all dietary miRNAs.

The model adopt a modified recursive feature elimination strategy with the grid parameter search:

First, the model randomly samples 60% data from positive instances and unlabeled instances (eg. dietary miRNAs) to train to manifold ranking with all features. Then, it adds the rest of 40% data as the unlabeled samples (includes the 40% positive data) to re-rank the entire dataset. As an evaluation, the model checks the final ranking list and count the number of true positive samples on the top of list. In this case, the model calculates the percentage of known circulating miRNAs that are ranked in the top 360 in the final ranking list. This process is conducted once for each parameter combination to infer the impacts of different parameter setting. Moreover, to avoid the probable bias from dataset, the model re-samples the training set for every run.

Secondly, according to the final feature ranking, which is aggregated from F-score, Fisher's ratio and Wilcoxon statistic *W*, the model halve the number of features each iteration, while keeping track of the top-ranked positive percentages for each round of feature elimination. By conducting this algorithm, the model obtains the interaction between feature size (along with the parameters) and the accuracy to find the most discriminative feature set and optimized parameters that produce the best predictive power.

Finally, with the selected features and optimized parameters, the model carries out a final manifold ranking to predict human absorbable exogenous miRNAs.

This ensemble manifold ranking model carefully considers the effects of parameter setting and the power of most discriminative features. It efficiently optimizes the predictive capabilities of traditional manifold ranking algorithm.

III. RESULTS

A. The features distinguish the transportable miRNAs from the rest

Consequently, the model selected 70 features to conduct the final manifold ranking. The selected features are categorized into five groups in TABLE III.

As expected, 63 selected features are related to the nucleotide composition of the sequences, such as single nucleotide C in the seed region, and tri-nucleotide AUA in the precursor sequence. As we mentioned above, the binding strength between miRNA and exosomes or AGO protein may plays a critical role on deciding if a miRNA can be transportable into human circulation or not, so those nucleotide compositions may reflects the impacts of this factor. Beside of the sequential features, some structural indicators are included into the final feature set as well. There are 5 frequencies of triplet nucleotide structures, such as C(((, A.(..The minimum free energy (MFE) also plays an important role in differentiating the circulation miRNA against others.

B. Transferrable exogenous miRNAs prediction

The ensemble manifold ranking model finally predicts the human absorbable exogenous miRNAs based on 70 selected features. As an important assessment of the ranking algorithm, 350 (~97%) of 360 human circulating miRNAs are ranked among top 360 in the final ranking list, which indicates the ranking models are well trained. Theoretically, any exogenous miRNA, which is ranked above a known blood miRNA, should be categorized as a transferrable exogenous miRNA

However, in order to minimize the possible false positive cases, we applied a strict rale to only consider the dietary miRNA transportable only if it has been ranked above all human circulating miRNA. Finally, 461 dietary miRNAs are predicted as human absorbable exogenous miRNAs.1

The 74 top-ranked transportable dietary miRNAs from prediction are shown in TABLE IV. The complete ranking list can be downloaded at http://go.unl.edu/ormw.

Validation of predicted transferrable miRNAs

To further assess our prediction, we conducted a in-house cow's milk consumption experiment. The blood samples were collected from five health adult participants at 4 time points (0, 3, 6, 9 hours) after they consumed 1-liter milk. The total RNA from the pooled blood samples for each time point is subject to a small RNA sequencing analysis by Illumina-HiSeq2000. For data analysis, CAP-miRSeq [34] was employed to calculate the microRNA expressions. The annotation from *miRBase* (version 21) [35] was used as the reference library when mapping the reads to known miRNA sequences. We have carefully filtered out the low quality reads and strictly aligned high quality reads to all known mature miRNA sequences, precursor sequences and the genomes of human and cow.

In total, we identified 22 cow's milk miRNAs in the human blood samples, and three highly predicted bovine's milk miRNAs (bta-mir-181b, −26b, −23a) are validated in this

experiment. This result confirms that predicted transferrable exogenous miRNAs could indeed be integrated into the human circulation.

It is also well documented that virus miRNAs have the capabilities of transporting into mammalian circulation and targeting genes in the host organism after the viral infection [36]. Thus, besides the dietary miRNAs, we also utilize the ensemble manifold ranking model to identify the transferrable miRNAs from two vims species: Epstein-Barr virus (EBV) and Rhesus lymphocrypto-virus (rLCV). 26 virus miRNAs (11 from EBV and 15 from rLCV miRNAs) were predicted as transportable exogenous miRNAs by using our ensemble manifold ranking model.

In 2012, Riley et al. discovered that Epstein-Barr virus is able to regulate human gene expression and transforms human B cells to maintain its viral latency [36]. They identified 44 EBV miRNAs and their human target genes in the EBV transformed B cells through the HITS-CLIP sequencing. As expected, all 11 predicted EBV miRNAs (ebv-mir-bartl4–3p, bart5–3p, bart5–5p, bart7–3p, bart7–5p, bartl4–5p, bart9–3p, bart8–3p, bart8–5p, bartl3–5p, bart19–3p) have been also identified in Riley's study. Similarly, all 15 predicted miRNAs from Rhesus lymphocryptovirus (rLCV) (rlcv-mir-rll-l-3p, rll-7–5p, rll-17–3p, rll-17–5p, rll-16–3p, rll-19–3p, rll-16–5p, rll-33–3p, rll-24–3p, rll-7–3p, rll-24–5p, rll-10–3p, rll-l-5p, rll-2–5p, rll-33–5p) that are highly transportable in our prediction have been reported in [37] where Raily et al. have found these rLCV miRNAs detectable in B cells of infected mammilla samples by using deep sequencing.

Above validation results again confirm that the predictive power of our ensemble manifold ranking model is trustworthy.

IV. Conclusions

In this paper, we demonstrated a well-designed ensemble manifold ranking model to identify the human absorbable exogenous miRNAs from 14 common food species. Different from the traditional ranking algorithms, this model integrates the feature selection capability and parameter optimization to maximize the predictive power. 1,120 sequential and structural features were extracted to distinguish the human circulating miRNAs. The result shows that 461 dietary miRNAs were predicted as human absorbable miRNAs based on 70 selected discriminative features. According to both internal and external validation experiments, evidences strongly support that the performance of our ensemble manifold ranking model is highly promising.

Acknowledgment

The authors would like to thank all the individuals who have participated in this study. In particular, we thank Drs. Scott Baier and Janos Zempleni at UNL for their technical assistance in preparing the validation samples for sequencing analysis. We also appreciate the UNL Holland Computing Center for providing the computational facility and offer helpful technical discussion with their stuffs (Drs. Jingchao Zhang and Adam Caprez). This work is funded by National Institutes of Health (NIH) through the Nebraska Center for the Prevention of Obesity Diseases through Dietary Molecules (NPOD) (P20GM104320). Computational experiments conducted at UNL Holland Computing Center.

References

[1]. Bartel DP, "MicroRNAs: target recognition and regulatory functions," Cell, vol. 136, no. 2, pp. 215–33, 1 23, 2009. [PubMed: 19167326]

[2]. Fabian MR, Sonenberg N, and Filipowicz W, "Regulation of mRNA translation and stability by microRNAs," Anmi Rev Biochem, vol. 79, pp. 351–79, 2010.

[3]. Kozomara A, and Griffiths-Jones S, "miRBase: annotating high confidence microRNAs using deep sequencing data," Nucleic Acids Res, vol. 42, no. Database issue, pp. D68–73, 1, 2014. [PubMed: 24275495]

[4]. Friedman RC, Farh KK, Burge CB, and Bartel DP, "Most mammalian mRNAs are conserved targets of microRNAs," Genome Res, vol. 19, no. 1, pp. 92–105, 1, 2009. [PubMed: 18955434]

[5]. Baier SR, Nguyen C, Xie F, Wood JR, and Zempleni J, "MicroRNAs are absorbed in biologically meaningful amounts from nutritionally relevant doses of cow milk and affect gene expression in peripheral blood mononuclear cells, HEK-293 kidney cell cultures, and mouse livers," J Nutr, vol. 144, no. 10, pp. 1495–500, 10, 2014. [PubMed: 25122645]

[6]. Zhang L, Hou D, Chen X, Li D, Zhu L, Zhang Y, Li J, Bian Z, Liang X, Cai X, Yin Y, Wang C, Zhang T, Zhu D, Zhang D, Xu J, Chen Q, Ba Y, Liu J, Wang Q, Chen J, Wang J, Wang M, Zhang Q, Zhang J, Zen K, and Zhang CY, "Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA," Cell Res, vol. 22, no. 1, pp. 107–26, 1, 2012. [PubMed: 21931358]

[7]. Valadi H, Ekstrom K, Bossios A, Sjostrand M, Lee JJ, and Lotvall JO, "Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells," Nat Cell Biol, vol. 9, no. 6, pp. 654–9, 6, 2007. [PubMed: 17486113]

[8]. Hunter MP, Ismail N, Zhang X, Aguda BD, Lee EJ, Yu L, Xiao T, Schafer J, Lee ML, Schmittgen TD, Nana-Sinkam SP, Jarjoura D, and Marsh CB, "Detection of microRNA expression in human peripheral blood microvesicles," PLoS One, vol. 3, no. 11, pp. e3694, 2008. [PubMed: 19002258]

[9]. Diehl P, Fricke A, Sander L, Stamm J, Bassler N, Htun N, Ziemann M, Helbing T, El-Osta A, Jowett JB, and Peter K, "Microparticles: major transport vehicles for distinct microRNAs in circulation," Cardicrvasc Res, vol. 93, no. 4, pp. 633–44, 3 15, 2012.

[10]. Turchinovich A, Weiz L, Langheinz A, and Burwinkel B, "Characterization of extracellular circulating microRNA," Nucleic Acids Res, vol. 39, no. 16, pp. 7223–33, 9 1,2011. [PubMed: 21609964]

[11]. Arroyo JD, Chevillet JR, Kroh EM, Ruf IK, Pritchard CC, Gibson DF, Mitchell PS, Bennett CF, Pogosova-Agadjanyan EL, Stirewalt DL, Tait JF, and Tewari M, "Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma," Proc Natl Acad Sci USA, vol. 108, no. 12, pp. 5003–8, 3 22, 2011. [PubMed: 21383194]

[12]. Rinck A, Preusse M, Laggerbauer B, Lickert FI, Engelhardt S, and Theis FJ, "The human transcriptome is enriched for miRNA-binding sites located in cooperativity-permitting distance," RNA Biol, vol. 10, no. 7, pp. 1125–35, 7, 2013. [PubMed: 23696004]

[13]. Kandeel M, Al-Taher A, Nakashima R, Sakaguchi T, Kandeel A, Nagaya Y, Kitamura Y, and Kitade Y, "Bioenergetics and gene silencing approaches for unraveling nucleotide recognition by the human EIF2C2/Ago2 PAZ domain," PLoSOne, vol. 9, no. 5, pp. e94538, 2014.

[14]. Zhou J, Cheng Y, Yin M, Yang E, Gong W, Liu C, Zheng X, Deng K, Ren Z, and Zhang Y, "Identification of novel miRNAs and miRNA expression profiling in wheat hybrid necrosis," PLoS One, vol. 10,no. 2, pp. e0117507, 2015. [PubMed: 25706289]

[15]. Jones-Rhoades MW, and Bartel DP, "Computational identification of plant microRNAs and their targets, including a stress-induced miRNA," Mol Cell, vol. 14, no. 6, pp. 787–99, 6 18, 2004. [PubMed: 15200956]

[16]. Zhan S, and Lukens L, "Identification of novel miRNAs and miRNA dependent developmental shifts of gene expression in Arabidopsis thaliana," PLoS One, vol. 5, no. 4, pp. el0157, 2010.

[17]. Campo S, Peris-Peris C, Sire C, Moreno AB, Donaire L, Zytnicki M, Notredame C, Llave C, and San Segundo B, "Identification of a novel microRNA (miRNA) from rice that targets an alternatively spliced transcript of the Nramp6 (Natural resistance-associated macrophage protein

6) gene involved in pathogen resistance," New Phytol, vol. 199, no. 1, pp. 212–27, 7, 2013. [PubMed: 23627500]

[18]. Maragkakis M, Vergoulis T, Alexiou P, Reczko M, Plomaritou K, Gousis M, Kourtis K, Koziris N, Dalamagas T, and Hatzigeorgiou AG, "DIANA-microT Web server upgrade supports Fly and Worm miRNA target prediction and bibliographic miRNA to disease association," Nucleic Acids Res, vol. 39, no. Web Server issue, pp. W145–8, 7, 2011. [PubMed: 21551220]

[19]. Mitra R, and Bandyopadhyay S, "MultiMiTar: a novel multi objective optimization based miRNA-target prediction method," PLoS One, vol. 6, no. 9, pp. e24583,2011. [PubMed: 21949731]

[20]. Oulas A, Karathanasis N, Louloupi A, Iliopoulos I, Kalantidis K, and Poirazi P, "A new microRNA target prediction tool identifies a novel interaction of a putative miRNA with CCND2," RNA Biol, vol. 9, no. 9, pp. 1196–207, 9, 2012. [PubMed: 22954617]

[21]. Chiang K, Shu J, Zempleni J, and Cui J, "Dietary MicroRNA Database (DMD): An Archive Database and Analytic Tool for Food-Borne microRNAs," PLoS One, vol. 10, no. 6, pp. e0128089, 2015. [PubMed: 26030752]

[22]. Weber JA, Baxter DH, Zhang S, Huang DY, Huang KH, Lee MJ, Galas DJ, and Wang K, "The microRNA spectrum in 12 body fluids," Clin Chem, vol. 56, no. 11, pp. 1733–41, 11, 2010. [PubMed: 20847327]

[23]. Mathelier A, and Carbone A, "Large scale chromosomal mapping of human microRNA structural clusters," Nucleic Acids Res, vol. 41, no. 8, pp. 4392–408, 4, 2013. [PubMed: 23444140]

[24]. Meister G, Landthaler M, Patkaniowska A, Dorsett Y, Teng G, and Tuschl T, "Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs," Mol Cell, vol. 15, no. 2, pp. 185–97, 7 23, 2004. [PubMed: 15260970]

[25]. Lin CC, Jiang W, Mitra R, Cheng F, Yu H, and Zhao Z, "Regulation rewiring analysis reveals mutual regulation between STAT1 and miR-155–5p in tumor immunosurveillance in seven major cancers," Sci Rep, vol. 5, pp. 12063, 2015. [PubMed: 26156524]

[26]. Khan SS, Madden MG, "A survey of recent trends in one class classification," Artif. Intell. Cogn. Sci, vol. 6206, pp. 188–197 2010.

[27]. Luiz ACPLFC Lorena HN, Lorena Ana C, "Filter Feature Selection for One-Class Classification," Journal of Intelligent & Robotic Systems, 2014.

[28]. W. J. Zhou D, Gretton A, Bousquet O, Scholkopf B, editors., "Ranking on Data Manifolds," Bradford Book, 2004.

[29]. Bindewald E, Cestaro A, Hesser J, Heiler M, and Tosatto SC, "MANIFOLD: protein fold recognition based on secondary structure, sequence similarity and enzyme classification," Protein Eng, vol. 16, no. 11, pp. 785–9, 11, 2003. [PubMed: 14631066]

[30]. He J, Li M, Zhang HJ, Tong H, and Zhang C, "Generalized manifold-ranking-based image retrieval," IEEE Trans Image Process, vol. 15, no. 10, pp. 3170–7, 10, 2006. [PubMed: 17022278]

[31]. Liu Q, Cui J, Yang Q, and Xu Y, "In-silico prediction of blood-secretory human proteins using a ranking algorithm," BMC Bioinformatics, vol. 11, pp. 250, 2010. [PubMed: 20465853]

[32]. Chih-Chung Chang C-JL, "LIBSVM : a library for support vector machines," ACM Transactions on Intelligent Systems and Technology, 2011.

[33]. Pihur V, Datta S, and Datta S, "RankAggreg, an R package for weighted rank aggregation," BMC Bioinformatics, vol. 10, pp. 62, 2009. [PubMed: 19228411]

[34]. Sun Z, Evans J, Bhagwate A, Middha S, Bockol M, Yan H, and Kocher JP, "CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data,"BMC Genomics, vol. 15, pp. 423, 2014. [PubMed: 24894665]

[35]. Dahiya N, Sherman-Baust CA, Wang TL, Davidson B, Shih Ie M, Zhang Y, Wood W, 3rd, Becker KG, and Morin PJ, "MicroRNA expression and identification of putative miRNA targets in ovarian cancer," PLoS One, vol. 3, no. 6, pp. e2436, 2008. [PubMed: 18560586]

[36]. Riley KJ, Rabinowitz GS, Yario TA, Luna JM, Darnell RB, and Steitz JA, "EBV and human microRNAs co-target oncogenic and apoptotic viral and human genes during latency," EMBO J, vol. 31, no. 9, pp. 2207–21, 5 2, 2012. [PubMed: 22473208]

[37]. Riley KJ, Rabinowitz GS, and Steitz JA, “Comprehensive analysis of Rhesus lymphocryptovirus microRNA expression,” J Virol, vol. 84, no. 10, pp. 5148–57, 5, 2010. [PubMed: 20219930]

TABLE I.

DETAILED STATISTICS OF MICRORNA DATASET

Category	Species	Mature miRNAs	Precursor miRNAs
Dietary Speices	Apple	203	202
	Banana	360	180
	Corn	309	166
	Grape	180	157
	Orange	61	57
	Rice	634	526
	Soybean	620	554
	Tomato	96	68
	Wheat	111	108
	Cow’s milk	243	245
	Cow’s fat	205	229
	Atlantic Salmon	498	371
	Chicken	994	740
	Pig	411	382
Virus Speices	Epstein-Barr virus	44	25
	Rhesus lymphocrypto-virus	68	36

TABLE II.

Detailed Feature Descriptions

	Feature Details	Counts
I[a]	Single Nucleotide Frequency	12[b]
	Pairwise Nucleotide Frequency	48[b]
	Triplet Nucleotide Frequency	192[b]
	Quadruplet Nucleotide Frequency	768[b]
	A + U Frequency	3[b]
	G + C Frequency	3[b]
	G + U Frequency	3[b]
	Number of Palindromes in Sequence	3[b]
	Length	3[b]
	Pairs of A-U in Premature microRNA	1
	Pairs of G-C in Premature microRNA	1
	Pairs of G-U in Premature microRNA	1
II[a]	Triplet nucleotide structures	32
	Minimum Free Energy, Normalized Minimum Free Energy, etc.	3
	Ensemble Free Energy, Normalized Ensemble Free Energy	2
	Stem-loop Statistics (e.g.: Average Stem Length, Maximum Stem Length, etc.)	25
	Minimum Free Energy Statistics (eg: mfe/unpaired nucleotides,etc.)	6
	Percentage of sequence composing of pairs.	1
	Frequency of Nucleotides that occur outside of UA, GU, GC pairs.	4
	Predicted shape type probability base on RNAshapes.	5
	RNAshapes statistics (e.g.: Shannon Entropy)	4

[a]I represents sequential features and II represents structural features;

[b]the total number of corresponding features on three type of sequences, namely seed sequences, mature sequences, and precursor sequences.

TABLE III.

SELECTED FEATURED DESCRIPTIONS

Feature groups	Counts	Feature lists
Nucleotide frequency in seed sequence	22	C, A, GC, GG, UUC, GUG, GAG, GAGA, AUUG, AUAG, UCUA, CGUG, CUUC, GCGG, GGCC, CAAC, CAUG, UGGC, UUGC, UAUA, CCGA, GCGC
Nucleotide frequency in mature miRNA	16	UAG, CUC, GCG, ACC, AACU, UUGU, AGCG, UGCC, AACG, UGAA, UAUC, AUCC, GCUU, UUAA, GCUC, ACGA
Nucleotide frequency in precursor sequence	25	CC, GU, AUA, ACU, UGC, AGG, GCAC, CCUA, CUCA, CGAU, UAGU, ACGU, GCCG, GUGU, CCGU, CGAC, AUAC, UUUG AUCC, GGUA, GGAA, AUCU, CGAG, AUAU, UUGG
Frequency of triplet nucleotide structures	5	C(((, A.(., U(.(, C(.(, U(..
Structure indicator	1	Minimum free energy (MFE)
Stems/Pairs	1	pairGU

TABLE IV.

74 HIGHLY RANKED DIETARY MIRNAS

Species	Transportable Dietary miRNAs
Apple	mdm-mir3627b, mirl56i, mir7121f, mirl71j, mirl67c
Corn	zma-mir395h, mir395j, mir395e, mir395h
Grape	vvi-mir395d, mir395m, mir395j, mir395i, mirl67c, mir3951
Rice	osa-mir395c
Soybean	gma-mir394f, mir4412, mirl71e, mir395c, mirl520h
Cow's milk	bta-mir-125b, mir-30b-5p, mir-22-5p, mir-184, mir-338, let-7g, mir-99a-5p, mir-22-3p, mir-99a-3p, mir-16a, mir-224, mir-409a, let-7e, mir-494, mir-181b
Cow's fat	bta-mir-124b, mir-654, mir-429, mir-34b, mir-412
Chicken	gga-mir-144, mir-200a, mir-1555, mir-365, mir-200a, let-7g, mir-124b, mir-1679, mir-1702, mir-133b, mir-3532, mir-20b, mir-144, mir-1744, mir-219a, mir-1617, mir-365, mir-1555, mir-23b, mir-19b, mir-103, mir-99a, mir-103, mir-455
Pig	ssc-mir-455, mir-133b, mir-96, mir-184, let-7g, let-7a, mir-708, mir-125b, mir-124a